Improving OCR quality of Historical Newspapers with Handwritten Text Recognition Models

Clematide, Simon ; Ströbel, Phillip

Originally published at:
Clematide, Simon; Ströbel, Phillip (2018). Improving OCR quality of Historical Newspapers with Handwritten Text Recognition Models. In: DARIAH-CH Workshop, Neuchâtel, 29 November 2018 - 30 November 2018, University of Neuchâtel.

CONTACT PERSONS
Simon Clematide, UZH
Marten Düring, C2DH
Maud Ehrmann, EPFL

impresso Media Monitoring of the Past

STAY CONNECTED
https://impresso-project.ch
info@impresso-project.ch
TWITTER @ImpressoProject

Nº 0001 LAUSANNE, LUXEMBOURG, ZURICH / THURSDAY, NOVEMBER 29TH, 2018 FREE / OPEN-SOURCE

Overall impresso pipeline

Main objective: enabling critical text mining to search, extract, process, link, and explore data from print media archives via a unified web interface

INVOLVES

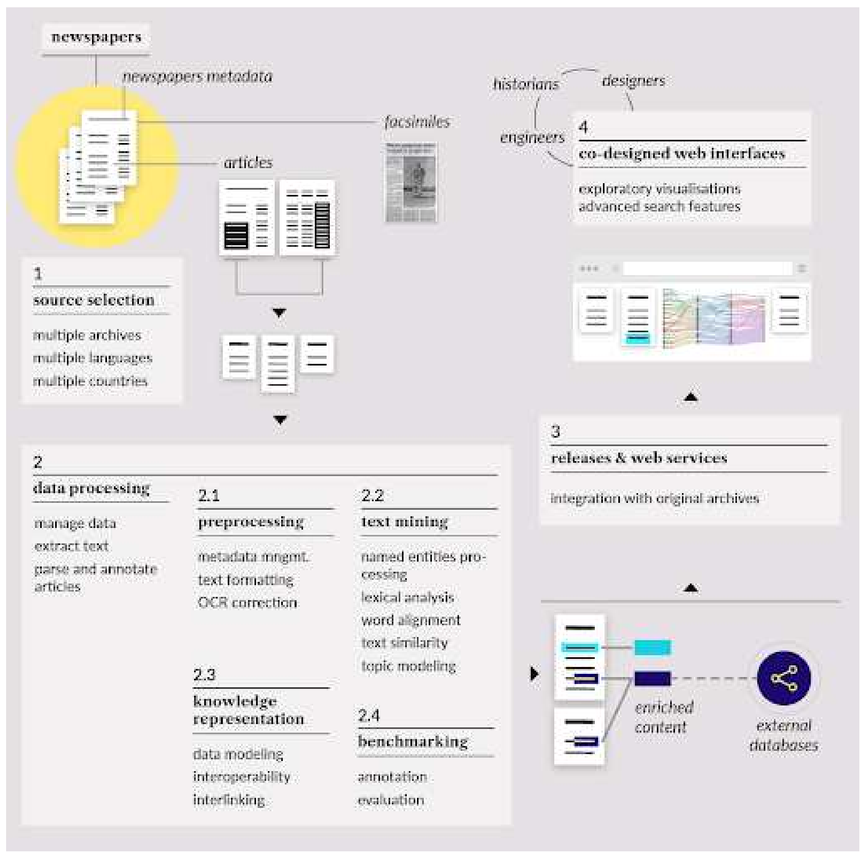
- computational linguists
- digital humanists
- historians
- web designers

THEIR COMMON GOAL

- tackling the challenges of content enrichment and data representation, visualization and analysis, completed by methodological and epistemological reflections

EXPECTED OUTCOMES

- natural language processing (NLP) tools dedicated to historical print media written in French and German
- visualization interfaces for active and goal-oriented exploration and critical analysis of newspaper corpora
- an application of digital history research on resistance to European integration.



Surprising success: Researchers report astonishing OCR results on historical newspapers!

OCR QUALITY OF HISTORICAL NEWSPAPERS

```
,»!’’ ‘.’’., ,!^ WWW’ ütchM ^!’’i: ...Me»
in l^t Ms«ndlM »edr»cks»n>; ‘!” ‘« ‘.IH,
! ‘,’ Großbrit«».««!.. .chOwtdf^.WnM««..
Hz<;ntschl«!»d,liN!ii.ung»Ntim Norden.^
Königes pou Isq.M’»«.^«!^^..^»««^«»’^
Prand »u., Debr«in). .Fl«nkl<;lch (Dü-
niouri« letzter Uuf»ll. Edle Vnft«lt
lu^ ««««ch). Gchwell WMatssM , ‘ G r Zu
lioZddn ‘«tstvtbe... «iß ,-.u»ch <;» ß
HZ l«,«‘» «il <; n.’ 5 welGs/ich EN).
s«ellen»’lü Mtlie». habt/ Iso« Hob Ihre«
«i»!«»«»..»..,’ di« Kapirin bewiest«»
«apfet. eit ihm unsre Newunderun« «wer-
ben h,«l”, ünd wege» des, bittet , stu
traMjlen M^etben« die «im». Vesttzü..«
mi«d bezeugen.- es <;3t«llt».^1MTheil-
nahck^ in Zw. Äbl^ gefiel prMMM ili H.
«r« L» »ar lmIaM.M.’«^«. AHstebdä.« MniH
«Mai»? b?«^ - >;«.ige«^ ? d«<; fi« «uf
.z. MM«,«« O«d«n «iNluft, erb. demÄll-
maWgtt. ihn ‘,u ffch/y» neumen/’ u«t»
ei lockmt i,s« stiiltl iu, wenn Auch
«icht bhn^HWintt/boHM’S^^ stin t»d»P-
tirtel<;S«h«, -Willia«. H<;l^>;e>;
```

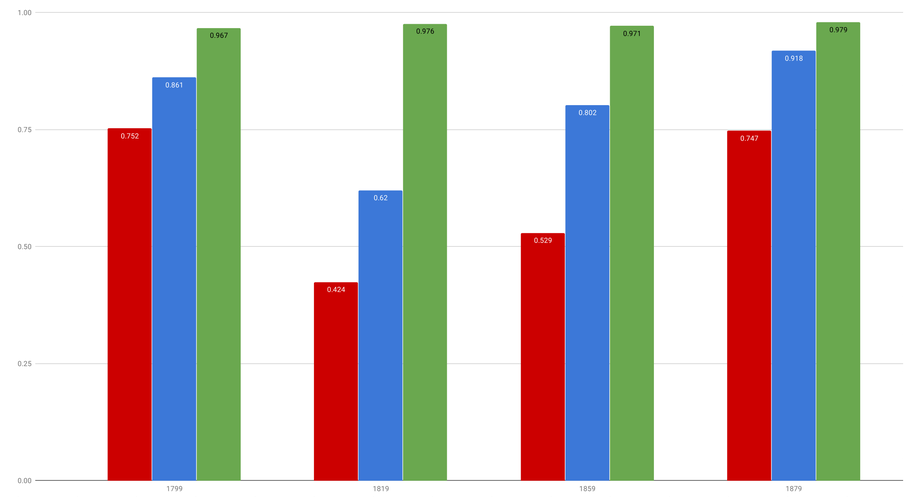
TACKLING OCR WITH HTR TOOLS

- 167 front pages from the Neue Zürcher Zeitung (NZZ)
- we used Transkribus[1] to create a gold standard
- manual correction of words and baselines
- training of Handwritten Text
- Recognition (HTR) model within
- Transkribus with 158 pages

1 https://www.transkribus.eu
2 https://www.abbyy.com

EVALUATION AND COMPARISON

- bag-of-words f-measure evaluation with TextEval 1.4
  - www.primaresearch.org/tools/PerformanceEvaluation
- compare three different outputs
  - original OCR by NZZ (ocr-2005)
  - re-OCRised material using ABBYY Finereader[2] (ocr-2017)
  - Transkribus’ HTR model (htr-2018)



DISCUSSION & OUTLOOK

- HTR models significantly increase OCR quality
- relatively small gold standards for training purposes suffice for decent OCR
- better OCR is beneficial for text mining techniques

Open questions:

- Do the HTR models trained on the NZZ perform equally well on other newspapers?
- How does occasionally occurring text in antiqua affect OCR quality?

SUPERVISORS Fréderic Kaplan, EPFL - Andreas Fickers, C2DH - Martin Volk, UZH
COLLABORATORS Thijs van Beek - Estelle Bunout, PhD - Daniele Guido, M.Sc. - Matteo Romanello, PhD - Paul Schroeder - Phillip Ströbel, M.A.

FNSNF Fonds national suisse Schweizerischer Nationalfonds Fondo nazionale svizzero Swiss National Science Foundation
grant n° CR- SII5_173719

EPFL École polytechnique fédérale de Lausanne

University of Zurich[UZH]

C2DH